Test S of Haplotype Concordance and Discordance

Mahnaz Khattak
Jinnah College for Women
University of Peshawar
Peshawar, Pakistan

Shuhrat Shah
Department of Statistics
University of Peshawar
Peshawar, Pakistan

Salahuddin
Department of Statistics
University of Peshawar
Peshawar, Pakistan

Abstract

The test presented here is based upon the proband and his/her affected as well as unaffected siblings. Here, the siblings are analyzed in terms of similarities of haplotypes. The proposed 'S' test is used in testing hypothesis that a particular disease has random pattern of inheritance against the alternative hypothesis that it has non-random pattern of inheritance. Probability distribution, mean and variance of the test are derived under the null hypothesis of random inheritance of the disease. It is then applied to data set of varying size of sibships having at least one affected and one unaffected sibs to investigate the existence of linkage disequilibrium.

1. Introduction

To demonstrate the heritability of a trait, one way is to provide evidence for its linkage with a known genetic marker in sib pair data, that is the two traits tend to be inherited together more often than would be expected by chance alone. This implies that the loci with alleles determining the two traits are located at the same chromosome with recombination frequency less than 0.5, that is they are linked.
Many workers have contributed a lot for the development and generalization of sib-pair methods in different situations.

The affected sib-pair (AS) methods assume the presence of a tightly linked disease susceptibility locus (DS) in the vicinity of HLA region. Sib pairs from different families are categorized according to whether they share none, one or two haplotypes identical by descent (IBD).

Sib pair method can be extended to include any relative that share at least one haplotype in common (Cantor & Roter, 1987; Cantor, 1988) and this could be generalized to apply to extended pedigrees (Elston & Stewart, 1988)

De Veries et al. (1976) used the criterion

F=(maximum –minimum number of haplotypes from one parent) + (maximum – minimum number of haplotypes from the other parent).

Green and Woodrow (1977) improved the test 'F' by using criterion 'N' which takes acount of the family size distribution as well. It is given by:

N = maximum haplotype frequency from one parent + (maximum haplotype frequency from the other parent.

Based on the affected sib pair data, Green et al. (1983) suggested another measure of association 'R' which uses sum of repeats of the haplotypes from both parents of the sibs.

A comparative study of these tests, their powers, relative merits and demerits, various extensions and generalizations is given by Green and Shah (1993). A new method for differentiating between groups of patients according to severity of disease proposed by Shah et al. (1995) is found to be an effective tool in analyzing data sets with respect to disease severity. Khattak et al. (2005) suggested a simpler and easier method to distinguish between recessive and dominant mode of disease inheritance, and also provides ways to estimate probability of the disease under consideration in the population.

2. The proposed 'S' test

The test presented here is based upon the proband and his/her affected as well as unaffected siblings. Here the siblings are analyzed in terms of similarities of haplotypes. The hypothesis to be tested is that the disease has random pattern of inheritance against the alternative hypothesis that it has non- random pattern. The test is stated as;

S = (sum of haplotypes from both parents in the affected sibs – the number of distinct haplotypes in the affected sibs) – (sum of haplotypes from both parents in the unaffected sibs – the number of distinct haplotypes in the unaffected sibs i.e. $S = (2m - k_1) - (2r - k_2)$.

Where m is the number of affected siblings, r is the number of unaffected siblings, k_1 and k_2 takes values 2, 3 or 4 as the siblings may share 2, 3 or 4 different genes in affected and non affected sibs. We assume here that the two parents are heterozygous.

If we proceed with m=2 and r=1, then the possible values of S will be calculated as in the following table.

Table 1: Possible combinations of haplotypes in affected and unaffected sibs with their S scores

m= 2	r=1	S_{21}	m=3	r=2	S_{31}	S_{32}
ac ac	ac	2	ac ac ac	ac ac	4	2
ac ad	ac	1	ac ac ad	ac ac	3	1
ac bc	ac	1	ac ac bc	ac ac	3	1
ac bd	ac	0	ac ac bd	ac ac	2	0
ac ac	ad	2	ac ac ac	ac ad	4	2
ac ad	ad	1	ac ac ad	ac ad	3	1
ac bc	ad	1	ac ac bc	ac ad	3	1
ac bd	ad	0	ac ac bd	ac ad	2	0
ac ac	bc	2	ac ac ac	ac bc	4	2
ac ad	bc	1	ac ac ad	ac bc	3	1
ac bc	bc	1	ac ac bc	ac bc	3	1
ac bd	bc	0	ac ac bd	ac bc	2	0
ac ac	bd	2	ac ac ac	ac bd	4	2
ac ad	bd	1	ac ac ad	ac bd	3	1
ac bc	bd	1	ac ac bc	ac bd	3	1
ac bd	bd	0	ac ac bd	ac bd	2	0

m=4				r=1	r=2		r=3			S41	S42	S43
ac	ac	ac	ac	ac	ac	ac	ac	ac	ac	6	4	2
ac	ac	ac	ad	ac	ac	ad	ac	ac	ad	5	3	1
ac	ac	ac	bc	ac	ac	bc	ac	ac	bc	5	3	1
ac	ac	ac	bd	ac	ac	bd	ac	ac	bd	4	2	0
ac	ac	ac	ac	ad	ac	ac	ac	ac	ac	6	5	3
ac	ac	ac	ad	ad	ac	ad	ac	ac	ad	5	4	2
ac	ac	ac	bc	ad	ac	bc	ac	ac	bc	5	4	2
ac	ac	ac	bd	ad	ac	bd	ac	ac	bd	4	3	1
ac	ac	ac	ac	bc	ac	ac	ac	ac	ac	6	5	3
ac	ac	ac	ad	bc	ac	ad	ac	ac	ad	5	4	2
ac	ac	ac	bc	bc	ac	bc	ac	ac	bc	5	4	2
ac	ac	ac	bd	bc	ac	bd	ac	ac	bd	4	3	1
ac	ac	ac	ac	bd	ac	ac	ac	ac	ac	6	6	4
ac	ac	ac	ad	bd	ac	ad	ac	ac	ad	5	5	3
ac	ac	ac	bc	bd	ac	bc	ac	ac	bc	5	5	3
ac	ac	ac	bd	bd	ac	bd	ac	ac	bd	4	4	2

S_{mr} indicates the S-score for 'm' affected and 'r' unaffected sibs in a sibship.

These are some possible combinations for variable values of m and r (m>r) and their respective S values where $S = (2m-k_1)- (2r-k_2)= (2m-r)-(k_1-k_2)$

Now for k_1 and k_2 taking values only 2,3, or 4, the variable 'S' gets only the following possible values:

2(m-r)-2 when $k_1 = 4$, $k_2 = 2$
2(m-r)-1 when $k_1 = 4$, $k_2 = 3$ or k_1 3, $k_2 = 2$
2(m-r) when k1= k2 =2 or 3 or 4
2(m-r) +1 when $k_1 = 2$, $k_2 = 3$ or $k_1 = 3$, $k_2 = 4$
2(m-r) +2 when $k_1 = 2$, $k_2 = 4$

The probability distribution of 'S' is shown in Table.2

Table 2: Probability distribution of 'S' for m+r sized sibships

S	P(S=s)
2(m-r)-2	$(1-2^{-m+1})^2(2^{-r+1})^2$
2(m-r)-1	$2(1-2^{-m+1})^2(2^{-r+1})(1-2^{-r+1})+2(2^{-m+1})(1-2^{-m+1})(2^{-r+1})^2$
2(m-r)	$(2^{-m+1})^2(2^{-r+1})^2 +4(2^{-m+1})(2^{-r+1})(1-2^{-m+1})(1-2^{-r+1})+(1-2^{-m+1})^2(1-2^{-r+1})^2$
2(m-r)+1	$2(2^{-m+1})^2(2^{-r+1})(1-2^{-r+1})+2(2^{-m+1})(1-2^{-m+1})(1-2^{-r+1})^2$
2(m-r)+2	$(2^{-m+1})^2(1-2^{-r+1})^2$

The sum of probabilities is equal to one, hence it shows that it is a complete probability distribution, and we can derive its mean and variance easily.
Derivation of mean and variance of 'S' under H_o

$$E(S) = \{2(m-r)-2\}\{(1-2^{-m+1})^2(2^{-r+1})^2 + \{2(m-r)-1\}\{2(1-2^{-m+1})^2(2^{-r+1})(1-2^{-r+1}) +2(2^{-m+1})(1-2^{-m+1})(2^{-r+1})^2 \{2(m-r)\}\{2(2^{-m+1})^2(2^{-r+1})^2 +4(2^{-m+1})(2^{-r+1})(1-2^{-m+1})(1-2^{-r+1}) + (1-2^{-m+1})^2(1-2^{-r+1})^2\} + \{2(m-r)+1\}\{2(2^{-m+1})^2(2^{-r+1})(1-2^{-r+1}) +2(2^{-m+1})(1-2^{-m+1})(1-2^{-r+1})^2\} + \{2(m-r)+2\}\{(2^{-m+1})^2(1-2^{-r+1})^2\} = 2(m-r) + 2^{-m+1}-2^{-r+1} = 2\{(m-r) +2^{-m} - 2^{-r}\}$$ and

$$E(S^2)= \{2(m-r)-2\}^2\{(1-2^{-m+1})^2(2^{-r+1})^2 + \{2(m-r)-1\}^2\{2(1-2^{-m+1})^2(2^{-r+1})(1-2^{-r+1})+2(2^{-m+1})(1-2^{-m+1})(2^{-r+1})^2 \{2(m-r)\}^2\{2(2^{-m+1})^2(2^{-r+1})^2 +4(2^{-m+1})(2^{-r+1})(1-2^{-m+1})(1-2^{-r+1}) + (1-2^{-m+1})^2(1-2^{-r+1})^2\} + \{2(m-r)+1\}^2\{2(2^{-m+1})^2(2^{-r+1})(1-2^{-r+1}) + 2(2^{-m+1})(1-2^{-m+1})(1-2^{-r+1})^2\} +\{2(m-r)+2\}^2\{(2^{-m+1})^2(1-2^{-r+1})^2\}$$

Hence,

$$\text{Var}(S) = E(S^2) - \{E(S)\}^2 = 2\{2^{-m+1}(1-2^{-m+1}) + 2^{-r+1}(1-2^{-r+1})\}$$

Thus mean and variance of S are;

$$2(m-r) + 2^{-m+}1-2^{-r+1} \text{ and } \quad 2\{2^{-m+1}(1-2^{-m+1}) + 2^{-r+1}(1-2^{-r+1})\} \text{ respectively.}$$

Now for any number of affected (m) and non-affected (r) in a sibship of size m+r, we can find the expected mean and variance of test S.

Table 3 shows various number of affected and unaffected sibs, using test S, along with their probabilities and expected means and variances. We have assumed the parents to be heterozygous i.e. they have no common haplotypes and the number of affected is greater than non- affected sibs i.e. m>r.

Table 3: Probability distribution of S for sibship of size m+r along with their means and variances

S_{21}	$P(S_{21})$	S_{31}	$P(S_{31})$	S_{32}	$P(S_{32})$	S_{41}	$P(S_{41})$
m=2,r=1		m=3,r=1		m=3,r=2		m=4,r=1	
0	¼	2	9/16	0	9/64	4	49/64
1	2/4	3	6/16	1	24/64	5	14/64
2	¼	4	1/6	2	22/64	6	1/64
				3	8/64		
				4	1/64		
Mean	1		5/2		3/2		17/4
Variance	1/2		3/8		7/8		7/32

(Continued)

S_{42}	$P(S_{42})$	S_{43}	$P(S_{43})$	S_{mr}	$P(S_{mr})$
2	49/256	0	49/1024	2(m-r)-2	$(1-2^{-m+1})^2(2^{-r+1})^2$
3	112/256	1	308/1024	2(m-r)-1	$(1-2^{-m+1})^2\ 2(2^{-r+1})\ (1-2^{-r+1})+2(2^{-m+1})$ $(1-2^{-m+1})(2^{-r+1})$
4	48/256	2	526/1024	2(m-r)	$(2^{-m+1})^2(2^{-r+1})^2\ +4(2^{-m+1})(2^{-r+1})\ (1-2^{-m+1})(1-2^{-r+1})+$ $(1-2^{-m+1})^2(1-2^{-r+1})^2$
5	16/256	3	132/1024	2(m-r)+1	$2(2^{-m+1})^2(2^{-r+1})\ (1-2^{-r+1})+2(2^{-m+1})\ (1-2^{-m+1})(1-2^{-r+1})^2$
6	1/256	4	9/1024	2(m-r)+2	$(2^{-m+1})^2(1-2^{-r+1})^2$
	13/4		7/4		$2\{(m-r)+2^{-m} - 2^{-r}\}$
	23/32		19/32		$2\{2^{-m+1}(1-2^{-m+1})+ 2^{-r+1}(1-2^{-r+1})\}$

3. Application to Data Sets

To illustrate the application of S we analyze data relating to rheumatoid arthritis which consists of 22 families with at least one member affected by the disease

and was collected by Cudworth and Woodrow (1975) found in the literature. The families were typed for HLA-A, B, Cw, and DR alleles by the hematology department. Table 4 present the data and S- scores as well as the expected means and variances for analysis.

Table 4: Cudworth's family data on rheumatoid arthritis and HLA typing:

family	affected sibs with haplotypes	unaffected sibs with haplotypes	S	E(S)	Var(S)
#1	a/b, c/b	c/b	1	1	1/2
#2	a/d, a/b	-	1	1	1/2
#3	a/b, a/b	-	2	1	1/2
#4	a/b, a/b	c/d	2	1	1/2
#5	a/b, a/d	-	1	1	1/2
#6	a/b, c/b	-	1	1	1/2
#7	a/b,a/d,a /b	-	3	5/2	3/8
#8	c/b, c/b	-	2	1	1/2
#9	a/b, c/d	-	0	1	1/2
#10	a/d, a/d	-	2	1	1/2
#11	a/b, c/b	-	1	1	1/2
#12	a/b,c/b,c/b	-	3	5/2	3/8
#13	a/b, a/b	a/d	2	1	1/2
#14	a/d, c/d	-	1	1	1/2
#15	a/d	a/d	0	0	1/2
#16	a/b	a/b	0	0	1/2
#17	a/d,c/d, a/b	c/d	2	5/2	0
#18	c/b,c/d,c/d	a/d, c/d	2	3/2	0
#19	c/b	-	0	0	3/8
#20	a/b, c/d	-	0	1	7/8
#21	a/b, a/b	-	2	1	0
#22	a/b	c/b	0	0	1/2
total			28	23	9

The test criterion yielded by S is;
$T = \{\sum S - \sum E(S)\}/ \{\sum Var(S)\}^{1/2}$

That is:

T = (28-.05-23)/3 = 1.667 & P= .0475 which is just significant at 5% level.

Hence it reveals some sort of linkage which is perhaps sufficient to suggest that the possibility of association between HLA and disease genes will be worth exploring when further samples are taken and of course the sample size be enlarged and the data be well randomized.

4. Conclusion

The test gives expected means and variances for any number of affected and non-affected sibs. If the number of unaffected are greater than the number of affected i.e. if for example we have S_{13} instead of S_{31} then the variance will remain same but mean will have same value with negative sign. Further for equal number of affected and unaffected sibs mean is always zero and variance is $4\{2^{-m+1}(1-2^{-m+1})\}$. The new test S is more simple and easily applicable then the already established tests for disease association with genes and their non-random inheritance.

The new test not only takes haplotype concordance among the affected sibs but it also considers haplotype discordance in the whole sibship. Usually the information given by non-affected sibs of the diseased person were ignored in the past. The new test S provides room for the incorporation of information contained in the unaffected sibs. The distribution of this test under H_o was the only way to calculate its mean and variance, and then apply it to data set for detecting linkage disequilibrium.

Reference

1. Cantor, R.M. (1988). A linkage test with identity by descent marker data from pairs of affected relatives. Prog. Clin. Bio. Res; 329: 111-116.
2. Cantor, R.M. and Rotter, J.I. (1987). Marker concordance in pairs of distant relatives: A new method of linkage analysis for common diseases. An. J. Hum. Genet. 39: 252-268.
3. De Veries, R.P.; Lai, A. and Fat, R.F.M. (1976). HLA-linked genetic control of host response to Mycobacterium Leprae. Lancet, 11, 1328-1330.
4. Elston, R.C. and Stewart, J. (1988). A general model for the genetic analysis of pedigree data. Hum. Hered, 21: 523-542.
5. Green, J. R. Woodrow, J.C. (1977). Sibling Method for detecting HLA linked genes and disease. Tissue antigens. 9:31-35.
6. Green, J. R., Low, H.C. and Woodrow, J. C. (1983). Inference on the inheritance of disease using repetitions of HLA haplotypes in affected siblings. Ann Hum Genet 47: 73-82.
7. Green, J.R. and Shah, S. (1993). Power comparison of various sibship tests of association. Ann. Hum. Genet. 57: 151-158.

8. Khattak, M., Shah, S. and Salahuddin. (2005). Mode of inheritance of HLA associated diseases. Pak. J. Statist. Vol. 21(2) 203-208.

9. Shah, S., Khattak, M. and ObaidulHaq, Qazi. (1995). A test of inheritance for disease severity. Pak. J. Hist. & Phil, Vol. 1: 29-36.